ISSN 2007-9737

Yuhmu Database: A Corpus of Tonal Speech Lacking Conventional Writing

Eric Ramos-Aguilar[1,2], J. Arturo Olvera-López[1,*], Ivan Olmos-Pineda[1]

[1] Benemérita Universidad Autónoma de Puebla, Facultad de Ciencias de la Computación, Puebla, México

[2] Instituto Politécnico Nacional, UPIIT, Tlaxcala, México

eric.ramosag@alumno.buap.mx, jose.olvera, ivan.olmos@correo.buap.mx

Abstract. This paper presents the development and analysis of a digital audio database of words pronounced in Yuhmu, a tonal and endangered variant of the Otomi language spoken in Ixtenco, Tlaxcala, Mexico. The database is composed of over 8,000 word recordings, including both correct and incorrect pronunciations, which were evaluated by native speakers through perceptual judgments. Statistical analyses reveal linguistic diversity in the phonetic components. Additionally, three experimental methodologies were implemented to evaluate the database: automatic segmentation of Mel-scale spectrograms using cosine distance, pronunciation classification via a multilayer perceptron, and implicit segmentation based on cosine distance thresholds. The results demonstrate good accuracy and successful detection of phonetic boundaries, which is comparable to methods applied to languages with a strong digital presence. This database constitutes a fundamental resource for the analysis of under documented tonal indigenous languages, highlighting the importance of preserving linguistic diversity. The controlled acoustic conditions and phonetic variability present in the database provide a solid foundation for future interdisciplinary studies in computational linguistics, machine learning, and language preservation.

Keywords. Low resource languages, Yuhmu Mexican language, phonetic representation of speech, database.

1 Introduction

The use of computational tools aimed at facilitating human work has increased throughout human development. These tools incorporate components capable of enabling communication between humans and machines. This interaction has seen significant advancements with the incorporation of new technologies, specifically artificial intelligence.

Computational processes supported by Artificial Intelligence have improved interaction between machines. As a result, machines today can autonomously or semi-autonomously perform various tasks, such as classifying, generating, and analyzing language; these tasks are typical of Natural Language Processing (NLP).

NLP is considered a subfield of Artificial Intelligence that aims to enable the automatic processing of human language through computational systems. This processing includes tasks such as understanding, generation, and structural transformation of language [9]. Implementing tasks within NLP requires the use of structured data that allows training Machine Learning (ML) models, which are capable of manipulating and generating a description of the relevant information.

As in other areas of Artificial Intelligence, NLP relies on sufficient data or instances to train models

for specific tasks. This raises important questions about the type of data required and its availability. Consequently, the quality and quantity of data become particularly relevant when comparing languages with sufficient digital presence to be processed in computational environments.

Currently, there are numerous models trained on high-resource languages such as English, Mandarin, or French, which benefit from structured data organized in well-defined formats like tables with rows and columns [4]. Access to high-quality linguistic data enables strong performance in various NLP tasks, such as automatic speech recognition, text generation, and machine translation. However, this advantage does not extend to languages with limited digital resources, commonly referred to as Low-Resource Languages (LRLs).

Analyzing these languages presents multiple challenges. To address them, approaches such as transfer learning and the creation of vocabulary embeddings have been explored to expand the available dataset. Nevertheless, these strategies are not always feasible, especially when there are significant linguistic differences or when the reference language has little similarity with the target language. In such cases, the expected outcomes are often limited.

One alternative is the creation of new linguistic corpora that allow training models without relying on a high-resource reference language. In many initiatives, the lack of a direct relationship between the source language and the target language limits the scope of the results compared to what can be achieved through linguistic transfer. Therefore, it becomes essential to develop digital resources from scratch for low-resource languages, such as Yuhmu, so that they can be computationally analyzed and processed.

This document presents a digital audio corpus for the Yuhmu language, a variant of Otomí spoken in Mexico, which has limited digital representation and is considered an LRL. This situation presents a significant challenge for the development of NLP technologies, as the lack of structured data hinders the training of accurate and functional models. The creation of this corpus seeks to address this scarcity through a localized and sustainable approach.

The construction of this corpus aims to lay the foundation for the computational analysis of Yuhmu, facilitating the development of technological tools that support its preservation, documentation, and potential revitalization. This effort responds not only to a computational need within the field of ML but also to a broader commitment to linguistic and cultural diversity.

2 Related Work

Currently, there has been limited work on the computational analysis of Indigenous languages of Mexico, mainly due to the multiple challenges posed by the lack of suitable digital databases for such tasks. The scarcity of structured resources significantly constrains the application of NLP and ML techniques to these languages.

Nevertheless, some repositories integrate audio and unstructured text data on Indigenous languages of Latin America, although many of them contain only between one and two thousand entries per language. A notable example is the Archive of the Indigenous Languages of Latin America (AILLA), a digital repository hosted by the University of Texas at Austin that preserves and disseminates multimedia materials in and about Latin American Indigenous languages, including audio recordings, texts, and videos. Another example is the UCLA Phonetics Lab Archive, a digital collection that brings together recordings of over 300 languages worldwide, accompanied by phonetic transcriptions and field notes. These resources have mainly been used in phonetic and phonological research and teaching, often from a more subjective and qualitative than computational perspective.

These repositories represent important advances in the preservation of Indigenous languages. However, there remains a significant need to generate corpora with structures and formats that enable their direct integration into automated linguistic analysis tasks.

Other studies, such as that developed by [7], address the generation of a digital audio corpus involving between five and approximately thirty speakers, with the goal of analyzing the Mixtec language through subjective analysis using Praat,

a free and open source software used for speech analysis and synthesis in phonetics [2].

Similarly, [13] describes the intonational characteristics of Hñöñhö (a variant of Otomi spoken in Tultepec, Queretaro, Mexico) through word segmentation using Praat, employing a database of recordings from 8 speakers, consisting of both interrogative and declarative sentences.

In the work of [16], machine translation systems were developed for several Indigenous languages of the Americas, including three from Mexico: Wixarika, Nahuatl, and Hñähñu. For Wixarika, a parallel corpus with Spanish based on the Zoquipan dialect was used, with translations provided by a native speaker, despite some inconsistencies in word boundaries.

In the case of Nahuatl, the training data were derived from an orthographically standardized version close to Classical Nahuatl, while the development and test sets were generated from translations into Modern Nahuatl, which were subsequently orthographically normalized. For Hñähñu, training texts were mostly from the dialect of the Mezquital Valley, whereas the evaluation data corresponded to the Ñûhmû dialect of Ixtenco, Tlaxcala. In all cases, multilingual pre trained models (such as mBART, Multilingual Bidirectional and Auto-Regressive Transformers) were fine tuned with these parallel datasets to address the challenges of translation learning in low resource settings.

As outlined above, the documentation of data on Indigenous languages of Mexico presents a significant challenge for current research, especially when no standardized reference for the target language exists. This lack of standardization complicates both data collection and systematic processing, limiting the applicability of computational approaches. In response to this situation, the development of a digital corpus for the Yuhmu language has been proposed, with the aim of facilitating its study from a computational perspective and enabling its integration into ML based applications. This corpus seeks to cover phonetic, phonological, and lexical aspects, and represents a valuable resource for advancing the automated analysis of low resource languages.

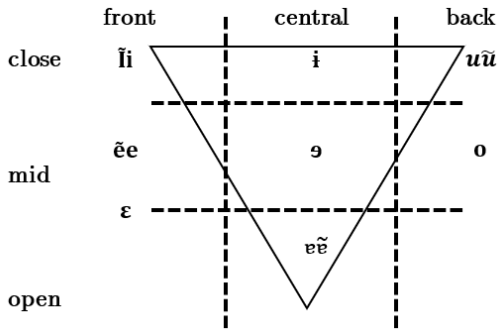

Fig. 1. Phonetic symbols of vowels (those with a tilde above them are considered nasal, while those without it are oral).

3 Yuhmu Language

The Yuhmu language is a tonal variety of Otomi, a macrolanguage belonging to the Otomangue linguistic group, spoken by an ethnic and cultural group settled in the south central region of Mexico.

Specifically, Yuhmu is spoken in the municipality of Ixtenco, Tlaxcala, Mexico. Currently, this language is critically endangered, as it is spoken by only a small number of elderly individuals (approximately over 70 years old) who still retain its pronunciation. Although some individuals under the age of 60 understand the language, there are no children acquiring it as a first language [3]. This situation has led to a marked decline in its use and intergenerational transmission, placing Yuhmu in the category of low resource languages due to its limited digital presence and the lack of structured data for analysis through computational tools.

According to a community census conducted by [3], it is estimated that there are around ± 75 Yuhmu speakers, although their level of linguistic competence has not been precisely documented. The language lacks its own writing system, which has led to various efforts to phonetically represent its sounds through the creation of isolated orthographic systems proposed by several historians.

Yuhmu has an inventory of 32 phonemes, which have been classified following the conventions of the International Phonetic Alphabet (IPA). Among these phonemes are 12 vowels (V), which can

Table 1. Symbols of the International Phonetic Alphabet for Consonants in Yuhmu.

Airway obstruction mode		Airway obstruction site				
Manner	**Voicing**	**Bilabial**	**Alveolar**	**Palatal**	**Velar**	**Glottal**
Plosive	Voiceless	p	t		k	k^w, ʔ
	Voiced	b	d		g	g^w
Affricate	Voiceless		ts	tʃ		
Fricative	Voiceless		s	ʃ		h
	Voiced		z			
Nasal	Voiced	m	n			
Tap or flap	Voiced		ɾ			
Approximant	Voiced			j		w

Fig. 2. Phases of the Yuhmu digital audio collection methodology.

occur in either oral or nasal variants, as illustrated in Fig. 1. In addition, 20 consonants (C) have been identified, grouped according to their place of articulation within the vocal tract. Unlike vowels, consonants are distinguished by their voicing, which can be either voiced or voiceless.

Table 1 presents a summary of the phonetic representation of Yuhmu consonants. The table is organized into two main columns, each describing specific phonetic properties:

Manner of airflow obstruction: This refers to how the airstream is modulated within the vocal tract during sound production. It includes categories such as plosives (where the airflow is momentarily stopped), affricates (a combination of stop and fricative), fricatives (where air passes through a narrow constriction causing friction), and nasals (where air passes through the nasal cavity), among others. It also classifies whether the vocal folds vibrate during production, distinguishing between voiced (with vibration) and voiceless (without vibration) sounds.

Place of airflow obstruction: This defines the exact point within the vocal tract where the airflow is constricted or blocked. Depending on the articulatory organs involved—such as the lips, tongue, or glottis—this obstruction can be classified into several types: bilabial (both lips), alveolar (alveolar ridge behind the upper teeth), palatal (hard palate region), velar (soft palate), or glottal (at the glottis or the space between the vocal folds).

Currently, the availability of documented digital data from native speakers is limited. Although recordings have been generated at some point, they are disorganized, lack labels, do not provide a defined structure, and the audio quality is often poor. This absence of digital information represents an obstacle for computational analysis and the development of Machine Learning (ML) models that could assist in analyzing the data.

Considering this context, a methodology has been designed to generate a database suitable for analysis in an artificial intelligence computational environment. This database aims to preserve the integrity and quality of the recorded information, ensuring that the data are a representative sample and suitable for training and evaluating ML models, thereby enabling a thorough and rigorous analysis of the Yuhmu language.

4 Proposed Methodology

This section describes the methodological process carried out for the creation of the Yuhmu language database, covering the fieldwork conducted, including the planning and selection of words, up to the validation of the obtained sample. The procedures used to capture audio recordings of native speakers are detailed, as well as the strategies implemented to ensure the quality and representativeness of the corpus. Fig. 2 shows the phases of the proposed methodology.

4.1 Population Identification

The fieldwork is based on the study conducted by [3], who estimates that there are approximately 75 active speakers of the Yuhmu language, distributed in the municipality of Ixtenco, Tlaxcala, whose total population is close to 15,000 inhabitants, according to INEGI. A preliminary exploration confirmed that most identified speakers are elderly adults, mostly over 80 years old, which complicates their location and participation. These native speakers represent the last generation to have acquired Yuhmu as their mother tongue. Additionally, some non native speakers were identified who learned the language in a family context, particularly through contact with grandparents or other direct relatives.

Given the context of linguistic vulnerability and considering the sensitive nature of approaching elderly individuals, a strategy based on pre-arranged interviews was designed. This allowed establishing trust relationships with the community, respecting the speakers' time, health conditions, and availability. Through the accompaniment of community mediators and local stakeholders, informed consent was obtained from each participant, thus ensuring an ethical and respectful data collection process.

Based on these conditions, it was decided to work with a sample of 24 Yuhmu speakers, which represents about 32% of the estimated population. This proportion is adequate and representative in studies of small linguistic communities, particularly in the context of Low Resource Languages [8].

4.2 Data Collection

The data collection process begins with a crucial preliminary stage that takes into account the work of [3], who documents the phonemes that comprise the Yuhmu language. Based on this reference, a foundational phonological dictionary was designed, consisting of 330 carefully selected words intended to include all 32 phonemes identified in the Yuhmu sound system. This dictionary not only ensures full coverage of the phonemic inventory but also encompasses various representative syllabic combinations of the language.

Although the aforementioned manual includes a broader set of terms, entries corresponding to complex grammatical constructions—such as combinations with articles or nominal phrases—were excluded in order to focus exclusively on simple lexical units that allow for clearer and more precise segmentation during the acoustic phonetic analysis.

This systematic approach ensures that the audio samples collected during fieldwork provide a balanced and comprehensive representation of the Yuhmu sounds, which is essential for training machine learning models. As an example of such representation, Table 2 provides a selection of words spoken in Yuhmu along with their phonetic transcription using the International Phonetic Alphabet (IPA).

Table 2 presents examples of five Spanish words and their corresponding pronunciation in Yuhmu, written using the International Phonetic Alphabet. Each entry provides the Spanish word (with its English translation in parentheses) and its phonetic transcription. The transcriptions reflect specific phonetic features of Yuhmu, such as glottalized consonants (e.g., /t'axi/ for "blanco") and aspirated

Table 2. Spanish words with phonetic representation of Yuhmu

Word in Spanish	Phonetic writing
abeja (bee)	/gáne/
barba (beard)	/kh̲úni/
blanco (white)	/t'axi/
calcetín (sock)	/nt'ókw'a/
hueso (bone)	/dó'yo/

sounds (e.g., /khúni/ for "barba"), also including the tonal features associated with the phoneme in the given word. This table serves to illustrate the diversity of documented phonemes and supports the design of acoustic models.

Once the base dictionary was selected, the recording phase proceeded using open source software. Audacity was chosen due to its widespread use, flexibility, and ease of use in fieldwork environments. Recordings were made using a BEHRINGER C1-U condenser microphone, which offers a frequency response range from 40 Hz to 20 kHz and a maximum sound pressure level of 136 dB (1% THD @ 1 kHz). This setup ensures a clear and accurate capture of the vocal signal, which is fundamental for subsequent acoustic analysis.

The recorded words were previously selected from the base dictionary, which is organized into semantic fields such as family, fruits, food, household and field elements, animals, body parts, among others. This classification not only ensures a representative coverage of the everyday Yuhmu lexicon but also a balanced distribution of phonemes in natural usage contexts.

All recordings were made in monophonic format (single channel) and stored as WAV files, which is an uncompressed standard preserving the original audio quality, a crucial aspect for subsequent phonetic acoustic and computational analyses [1].

This choice responds to the need to facilitate analysis in acoustically controlled environments, since stereo recordings (two channels) are usually optimized to enhance intelligibility in open spaces but may introduce undesired variability in closed settings [6].

All recording sessions were conducted in the homes of the interviewed speakers, always striving for the most controlled environment possible to minimize background noise, aiming to obtain high quality acoustic samples that allow for more precise phonetic and computational analysis.

4.3 Data Pre-processing

During the recording process, conversations were held with both native and non native speakers of Yuhmu, allowing the collection of a diverse set of linguistic data. However, this dynamic resulted in the inclusion of words that are not directly relevant to the specific objectives of the research. Additionally, some audio recordings exhibited background noise characterized by low amplitude and volume levels, reflecting acoustic conditions similar to typical environmental noise in natural settings.

Furthermore, recordings containing full sentences also showed a certain level of background noise, despite being conducted in controlled environments. This situation necessitates the application of digital audio segmentation techniques to isolate individual words within the corpus. This approach facilitates the removal of irrelevant elements, thereby optimizing the quality and precision of the pronunciation analysis within the generated corpus.

The collected sample includes approximately 6000 correctly pronounced words and around 2700 incorrectly pronounced words, distributed among the 24 interviewed speakers, none of which had undergone prior computational preprocessing. Of these speakers, 12 are native speakers, 7 are non-native with good pronunciation, and 5 are non-native with deficient pronunciation. During the review of the audio segments, background noise presence was identified, prompting the application of an attenuation process to reduce its impact.

The digital audio signal was segmented into specific words of interest. In some cases, the signal was amplified by 12 dB, nearly doubling the auditory power, which allowed for a notable increase in volume and improved listening quality. However, since amplification also increases the presence of unwanted signals, such as electrical

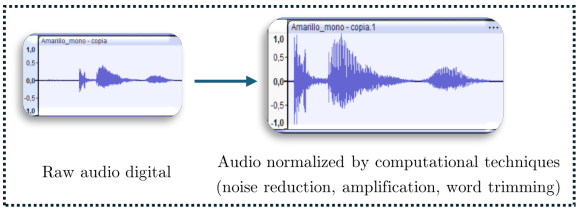

Fig. 3. Normalization representation of trimmed words

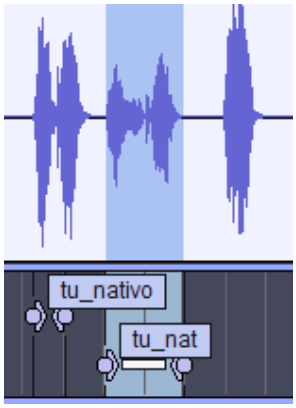

Fig. 4. Word clipping and labeling representation

noise, a second attenuation was sometimes applied to minimize these adverse effects (Fig. 3).

During this process, each word was carefully labeled according to its meaning in Spanish, which facilitated its organization and subsequent analysis. Finally, the labeled audio files were imported and systematically stored for use in computational analyses (Fig. 4).

A total of 5,835 words with correct pronunciation and 2,620 with incorrect pronunciation were obtained. The duration of the recordings ranged from 376 ms to 1.118 s, reflecting the variable length of the words depending on their phonemic complexity, which can range from a single phoneme to eleven phonemes per word.

The validation of correct pronunciation of the words was carried out through direct evaluation by native speakers. These evaluators, although lacking formal training in phonetics or graphical representations of Spanish, make a judgment based on auditory perception and their natural experience with the language. Through repetition and comparison of the words pronounced in Yuhmu, the native speakers issue a value judgment on whether the pronunciation is adequate or not.

It is important to emphasize that this judgment does not imply a blind validation of any sound production made by a learner; rather, the native speakers, intuitively, consider whether the phonemes have been articulated correctly. This occurs even though they lack technical knowledge of phonology, as their evaluation is based on a natural recognition of the phonetic patterns they have internalized since birth.

This validation criterion is fundamental, as it ensures that the pronunciation considered "correct" corresponds to the genuine and habitual linguistic practices of native speakers, thereby guaranteeing the quality and fidelity of the data for subsequent analyses.

4.4 Statistical Analysis

In order to better understand the scope, diversity, and representativeness of the speech corpus used in this study, a statistical analysis was conducted.

These calculations allow quantifying the volume and variability of the phonetic data, providing an essential perspective on the quality and robustness of the dataset with the aim of subsequently developing machine learning techniques.

It is important to mention that the sample is considered representative given the number of recorded individuals relative to the total population and the broad coverage of recorded words. Moreover, the data with correct pronunciation gain relevance since each word representation considers a number of phonemes with similarities among them, which reinforces its usefulness in the analysis.

This analysis was performed specifically on the database developed in the present work, which was constructed from a field collection process with native speakers and carefully organized to ensure its suitability for speech processing tasks. This database includes both correct and incorrect pronunciations and covers a wide range of phonetic combinations that allow for comparative studies and controlled experiments. The structure

and diversity of the corpus were key factors in defining the statistical criteria described herein.

The following data are given:

— Total number of correctly pronounced words: $N = 5835$

— Base dictionary: 330 words

— Each word is assumed to contain a number of phonemes that follows a discrete uniform distribution over the set:

$$\{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11\}$$

That is, the probability that a word contains k phonemes is given by:

$$P(\text{phonemes} = k) = \frac{1}{11}, \quad \text{for } k = 1, 2, \ldots, 11.$$

Let X be the random variable representing the number of phonemes per word, which follows a discrete uniform distribution over the values $1, 2, \ldots, 11$.

The expected value (mean) of X is:

$$E[X] = \frac{1}{11} \sum_{k=1}^{11} k = \frac{1}{11} \cdot \frac{11 \cdot (11+1)}{2} = \frac{11 \times 12}{2 \times 11} = 6.$$

The average number of phonemes per word is 6. The standard deviation of X is:

$$\sigma = \sqrt{E[X^2] - (E[X])^2},$$

where

$$E[X^2] = \frac{1}{11} \sum_{k=1}^{11} k^2 = \frac{1}{11} \cdot \frac{11 \cdot 12 \cdot 23}{6} = \frac{12 \cdot 23}{6} = 46.$$

Thus,

$$\sigma = \sqrt{46 - 6^2} = \sqrt{46 - 36} = \sqrt{10} \approx 3.16.$$

Multiplying the total number of words by the average number of phonemes per word:

$$\text{Total number of phonemes} \approx N \times E[X] =$$

$$5835 \times 6 = 35010.$$

These estimates are not merely illustrative but justify the relevance of the phonetic analysis strategies developed in this work. By knowing the volume, distribution, and diversity of the data, a solid quantitative basis is guaranteed, which supports both the choice of methodologies and the interpretation of the obtained results.

From a mathematical standpoint, the expected values and standard deviation obtained allow formal characterization of the behavior of the random variable X, corresponding to the number of phonemes per word. This, in turn, facilitates robust estimates regarding the total amount of phonetic information contained in the database. For instance, assuming a discrete uniform distribution, a mathematical expectation of 6 phonemes per word is obtained, allowing for a rigorous estimate that the 5,835 correctly pronounced words contain approximately 35,010 phonemes. This figure supports the use of statistical metrics and machine learning over a dataset sufficiently dense and diverse in terms of minimal sound units. Thus, the analysis not only accounts for the corpus size but also validates its suitability for segmentation and phonetic classification tasks.

Furthermore, this statistical analysis allows for more precise identification of particular challenges within the Yuhmu corpus, such as inter-speaker variability, word length, and the frequency of certain phonological structures. This information is key to designing more accurate tools for tasks such as automatic segmentation, pronunciation error detection, or the development of teaching materials aimed at speakers or learners. Therefore, the prior quantification of the data not only supports the methods employed but also guides future research directions based on this database.

5 Sample Evaluation

The use of ML models has been essential for applying various forms of analysis to the Yuhmu language. In particular, neural networks such as the Multilayer Perceptron (MLP) were employed, along with computational techniques based on cosine distance applied to spectrograms generated through systematic parameter searches. These tools enabled the implementation of classification

Table 3. Summary of three approaches applied to the phonetic analysis of Yuhmu (Ixtenco Otomi).

Description of the work	Main technique	Data used	Key result
Classification of correct and incorrect pronunciation using temporal and spectral features with MLP [12].	Feature extraction + MLP (Grid Search)	622 audios per class (correct/incorrect), based on 330 words	Accuracy of 90–97.7%; best results with windows of 20–40 ms.
Phonetic segmentation using Mel spectrograms and cosine distance to detect boundaries between phonemes [11].	Explicit segmentation + spectral analysis	297 recordings with complete phonemic combinations	SER between 23.89% and 26.03%; best performance with 20 ms window, 25% overlap, and 35–45 Mel filters.
Implicit segmentation and generation of phonetic images through thresholding applied to cosine distance matrices [10].	Implicit segmentation + visual analysis (MSE, SAM, SCC)	300 recordings, 66,559 representations, and 95,040 images	SER of 0% in over 18,000 combinations; high visual consistency among segments.

tasks and the evaluation of results using specific metrics such as Segmentation Error Rate (SER) and Accuracy.

Additionally, statistical analyses were conducted using measures such as Mean Squared Error (MSE), Spectral Angle Mapper (SAM), and Spearman's Correlation Coefficient (SCC), all of which are relevant for assessing the quality of segmentation and classification processes, especially in the context of phonetic and tonal analysis of low-resource languages like Yuhmu. These indicators were used to verify how similar the resulting phonetic images are across different samples or experimental conditions. These approaches have proven useful for validating both the database used and the implemented methodologies.

Table 3 summarizes three studies aimed at analyzing the Yuhmu language from different phonetic perspectives, each employing distinct methodologies. The first study incorporates both temporal and Mel-scale spectral features, using a multilayer perceptron with various hyperparameter configurations to classify phonemes and identify the most effective spectral combination [12]. The second study applies a methodology based on identifying the optimal spectrogram for phoneme segmentation, using cosine distance between columns and peak selection to highlight relevant acoustic information. This approach includes an explicit evaluation of the phonemes (manual), without the use of machine learning techniques, and is assessed using the Segment Error Rate (SER) metric [11]. Despite the manual nature of the evaluation, the analyzed data yield favorable results. Building on this, the third study investigates an implicit segmentation strategy through statistical analysis of an automatic phoneme segmentation method. This work focuses on identifying the phonetic patterns of Yuhmu and generating phonetic representations of its phonemes, proposing these representations as a solid foundation for building a spectral database. The results also show good consistency among the phonetic representations obtained, which have been labeled and validated [10].

These studies expand on research done with majority languages and show results that align well with previous findings [17, 14, 15, 5]. The data sample captures the full range of sounds and structures of the Yuhmu language, providing a realistic picture of how the language is spoken. To make sure the database is both high-quality and useful for studying under-resourced indigenous languages, various processing and machine learning methods were applied.

The consistency of results across different experiments shows just how robust the database is, even when dealing with the challenges of an endangered and little-documented language. Including phonetic variations and carefully controlling recording conditions makes the models developed from this data more reliable and generalizable. Beyond serving as a tool for linguistic analysis, this database has real-world potential: it could support speech recognition systems and educational resources designed specifically for the Yuhmu-speaking community.

6 Conclusions

The collected database, together with the analysis of previous studies on the Yuhmu language, constitutes a fundamental resource for the development of research based on machine learning techniques. This dataset contributes to academic knowledge and significantly supports the preservation of a language at risk of extinction.

The analysis of these data allows for the design of experiments that lay the groundwork for future research, including phonetic, tonal, and syntactic studies of a language that, in addition to being endangered, lacks a standardized writing system. This highlights the importance of continuing the systematic collection of languages whose sole representation is oral, emphasizing the need to preserve not only the sounds but also the cultural knowledge they convey.

The generated database represents an unprecedented contribution to the field of natural language processing for indigenous languages. It includes recordings that are structured, annotated, and validated by native speakers, encompassing variability in pronunciations, phoneme diversity, and controlled acoustic conditions. This enables robust experiments in classification, segmentation, and pronunciation assessment, providing a solid foundation for current studies and future interdisciplinary research involving computational linguistics, machine learning, and documentation of indigenous languages.

It is noteworthy that the database corresponds to a unique variant of Yuhmu, geographically isolated from other Otomi dialects in central Mexico. This characteristic gives it special linguistic value, as it allows the analysis of particular phonetic and morphosyntactic patterns not found in other regions, thereby contributing to the preservation of a specific and little-documented manifestation of this endangered language.

References

1. **Belodedov, M. V., Fonkants, R. V., Safin, R. R. (2023).** Development of an algorithm for optimal encoding of wav files using genetic algorithms. 2023 5th International Youth Conference on Radio Electronics, Electrical and Power Engineering (REEPE), IEEE, Vol. 5, pp. 1–6.

2. **Boersma, P., Weenink, D. (2025).** Praat: doing phonetics by computer [computer program].

3. **Esperanza, M. I. Y. G., Alarcón Montero, R. (2024).** Manual para la escritura de los sonidos del Yuhmu. Secretaría de Cultura, Instituto Nacional de Antropología e Historia, 1 edition.

4. **Hernández, C., Rodríguez, J. E. R., et al. (2008).** Preprocesamiento de datos estructurados. Revista vínculos, Vol. 4, No. 2, pp. 27–48.

5. **Lin, B., Wang, L. (2022).** Phoneme mispronunciation detection by jointly learning to align. ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, pp. 6822–6826.

6. **Liu, R., Zhang, J., Gao, G. (2024).** Multi-space channel representation learning for mono-to-binaural conversion based audio deepfake detection. Information Fusion, Vol. 105, pp. 102257.

7. **Penner, K. (2019).** Prosodic structure in Ixtayutla Mixtec: Evidence for the foot. Ph.D. thesis. DOI: 10.13140/RG.2.2.28786.96965.

8. **Petersen, L., Minkkinen, P., Esbensen, K. H. (2005).** Representative sampling for reliable data analysis: theory of sampling. Chemometrics and intelligent laboratory systems, Vol. 77, No. 1-2, pp. 261–277.

9. **Qiu, X., Sun, T., Xu, Y., Shao, Y., Dai, N., Huang, X. (2020).** Pre-trained models for natural language processing: A survey. Science China technological sciences, Vol. 63, No. 10, pp. 1872–1897.

10. **Ramos-Aguilar, E., Olvera-López, J. A., Olmos-Pineda, I. (2025).** Phonetic spectral image representation for yuhmu language analysis. Mexican Conference on Pattern Recognition, Springer, pp. 154–164.

11. **Ramos-Aguilar, E., Olvera-López, J. A., Olmos-Pineda, I., Sánchez-Rinza, B. E., Ramos-Aguilar, R. (2024).** Phonetic segmentation of the yuhmu language using mel-scale spectral representations. , pp. 29–40.

12. **Ramos-Aguilar, E., Olvera-López, J. A., Olmos-Pineda, I., Martín-Ortiz, M. (2023).** A general overview of language pronunciation analysis based on machine learning. Vol. 152, No. 10, pp. 29–43.

13. **Velásquez Upegui, E. P. (2020).** Entonación del español en contacto con el otomí de san ildefonso tultepec: enunciados declarativos e interrogativos absolutos. Anuario de letras. Lingüística y filología, Vol. 8, No. 2, pp. 143–168.

14. **Ye, W., Mao, S., Soong, F., Wu, W., Xia, Y., Tien, J., Wu, Z. (2022).** An approach to mispronunciation detection and diagnosis with acoustic, phonetic and linguistic (apl) embeddings. ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, pp. 6827–6831.

15. **Zhang, Z., Wang, Y., Yang, J. (2022).** Masked acoustic unit for mispronunciation detection and correction. ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, pp. 6832–6836.

16. **Zheng, F., Reid, M., Marrese-Taylor, E., Matsuo, Y. (2021).** Low-resource machine translation using cross-lingual language model pretraining. Proceedings of the First Workshop on NLP for Indigenous Languages of the Americas, Association for Computational Linguistics, pp. 104–111.

17. **Zou, W., Jiang, D., Zhao, S., Yang, G., Li, X. (2018).** Comparable study of modeling units for end-to-end mandarin speech recognition. 2018 11th International Symposium on Chinese Spoken Language Processing (ISCSLP), IEEE, pp. 369–373.

Article received on 30/05/2025; accepted on 10/07/2025.
**Corresponding author is J. Arturo Olvera-Lóopez.*